2000

Cross-Language Information Retrieval using Dutch Query Translation

Anne R. Diekema
Syracuse University

Wen-Yuan Hsiao
Syracuse University

Cross-Language Information Retrieval using Dutch Query Translation

Anne R. Diekema and Wen-Yuan Hsiao

Syracuse University
School of Information Studies
4-206 Ctr. for Science and Technology
Syracuse, NY 13244-4500, USA
{diekemar, whsiao}@syr.edu

Abstract. This paper describes an elementary bilingual information retrieval experiment. The experiment takes Dutch topics to retrieve relevant English documents using Microsoft SQL Server version 7.0. In order to cross the language barrier between query and document, the researchers use query translation by means of a machine-readable dictionary. The Dutch run was void of the typical natural language processing techniques such as parsing, stemming, or part of speech tagging. A monolingual run was carried out for comparison purposes. Due to limitations in time, retrieval system, translation method, and test collection, there is only a preliminary analysis of the results.

Introduction and problem description

Cross-Language Information Retrieval (CLIR) systems enable users to formulate queries in their native language to retrieve documents in foreign languages [1]. In CLIR, retrieval is not restricted to the query language. Rather queries in one language are used to retrieve documents in multiple languages. Because queries and documents in CLIR do not necessarily share the same language, translation is needed before matching can take place. This translation step tends to cause a reduction in cross-language retrieval performance as compared to monolingual information retrieval. The literature explores four different translation options: translating queries (e.g. [2], [3]), translating documents [4], [5], translating both queries and documents [6], and cognate matching [1] [7]. The prevailing CLIR approach is query translation.

The translation of queries is inherently difficult due to the lack of a one-to-one mapping of a lexical item and its meaning. This creates lexical ambiguity. Further, query translation is complicated by the cultural differences between language communities and the way they lexicalize the world around them. These two translation issues create many different translation problems such as lexical ambiguity, lexical mismatches, and lexical holes. In turn, these and other translation problems result in translation errors which impact CLIR retrieval performance.

The Cross-Language Evaluation Forum (CLEF) provides a multilingual test collection to study CLIR using European languages. One of the CLEF tasks is bilingual information retrieval. The aim of the bilingual task is the retrieval of documents in a language different from the topic (query) language. Unlike the multilingual task, only two languages are involved and retrieval results are monolingual. For the bilingual run we used the Dutch topic set (40 topics) to retrieve English documents (Los Angeles Times of 1994 – 113,005 documents, 409,600 KB). We were completely oblivious to CLEF and its deadlines but we happened to hear that CLEF results were due in one week. We immediately signed up and started on our mad rush to get results in on time.

Experimental Setup

In monolingual information retrieval experiments, researchers commonly vary the information retrieval system while keeping the test queries and documents constant. This allows for comparison between systems and comparison between different versions of the same system. The same practice is followed

[1] Cognate matching facilitates matching cognates (words that have identical spelling) across languages by allowing for minor spelling differences between the cognates.

in CLIR experiments when comparing different systems. However, CLIR experiments vary the test queries rather than the system, to allow for comparison between the cross-language and monolingual capabilities of the same system. The experiments in this research rely on varying the test queries.

By manually translating test queries into a foreign language and using these test queries as the cross-language equivalents, the cross-language performance of a system can be compared directly to its monolingual performance (see figure 1). Manual translation of queries is now a widely used evaluation strategy because it permits existing test collections to be inexpensively extended to any language pair for which translation resources are available. The disadvantage of this evaluation technique is that manual translation requires the application of human judgment, and evaluation collections constructed this way exhibit some variability based on the terminology chosen by a particular translator.

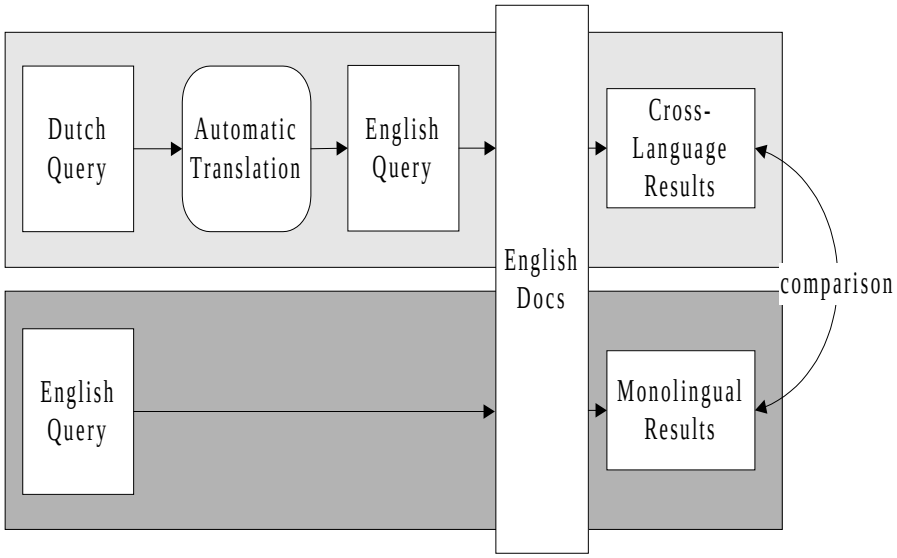

Fig. 1. Bilingual CLIR system evaluation.

The CLEF experiments described in this paper are modeled after the experiments described above. CLEF provided topic sets in both languages. Of these, we used only the descriptions and narratives. The English topics were pos-tagged to aid phrase detection and stopwords were filtered out using the SMART stop list. We wrote a crude perl program to convert the English query into a Boolean representation that was usable by the retrieval system (described in experimental setup). The Dutch topics were processed differently since we lacked Dutch text processing resources. For each query, we extracted individual tokens, treating each token separated by spaces as a single word. A dictionary lookup took place for each token and all possible translations with their parts of speech (nouns, adjectives, verbs, and adverbs only) were inserted into the query translation file. Words that lacked a translation were left untranslated. The translation file was converted into a logical representation. Translation synonyms were combined using the OR operator and phrases were added using double quotes around the phrase. We assumed that capitalized translated tokens were important to the query and used the AND operator to add them to the logical representation (see table 1).

Unfortunately our plain and simple approach was thwarted by the retrieval system which stumbled on our rather lengthy query representations. Since we only had hours to spare before we had to submit our results, we decided to drastically shorten our Dutch queries. The translations we used were grouped by part-of-speech so we decided to pick only those translations listed under the very first part-of-speech. The queries were still too long so we further limited the translation to the first term within that part-of-speech (excluding all synonyms). Looking back, we should probably have limited our queries to the title fields rather than using the lengthy description and narrative but we ran out of time. It is not surprising that our results were a bit dismal (see *Results*).

Original topic
<top> <num> C034 <D-title> Alcoholgebruik in Europa <D-desc> Omvang van en redenen voor het gebruik van alcohol in Europa. <D-narr> Behalve algemene informatie over het gebruik van alcohol in Europa is ook – maar niet uitsluitend – informatie over alcoholmisbruik van belang. </top>
Logical representation after translation (based on description and narrative)
("Europe") AND ("alcoholgebruik" OR "dimension" OR "application" OR "alcohol" OR "general" OR "data" OR "exclusively" OR "advantage")

Table 1. Query proces sing.

System Overview

The system used in the experiments utilized the full-text support of Microsoft SQL Server version 7.0 [8]. SQL Server is a commercial relational database system. Besides regular relational operations, in version 7.0, it introduces facilities that allow full text indexing and searching of textual data residing in the server. Full-text search on database data is enabled by proprietary extensions to the SQL language. The following search methods are available in SQL Server 7.0:

- search on words or phrases
- search based on prefix of a word or phrase
- search based on word or phrase proximity
- search based on inflectional form of verb or adjective
- search based on weight assigned to a set of words or phrases

However, we only used the phrase and word or phrase proximity search functions in the experiments described in this paper. The system requires documents in the collection to be exported to the database before any indexing and searching can take place. Therefore, a table was created in SQL Server to represent the whole collection and each document in the collection was converted to a record in the table. The table was comprised of two columns: DOCNO and DOCTEXT. DOCNO served as the unique identification of each record in the table. DOCTEXT stored the text content of the documents. In the TREC collection, all documents are marked up in standard generalized mark up language (SGML) format. Elements like DOCNO, TITLE, AUTHOR, and TEXT for example, are used to mark up text segments and to indicate the semantics of that portion of text. Among those elements, text content of each document's DOCNO element and the TEXT element was extracted and written into the table's DOCNO and DOCTEXT columns respectively. Any SGML tags inside the TEXT elements were stripped out before the actual export took place. After the table was populated with textual data from the collection, a full-text index was created based on the table's DOCTEXT column.

After a query was sent to the system, a result set of document number, DOCNO, along with rank was returned. The rank was a value between 0 and 1000 which was generated by SQL Server to indicate how well a record matched the query. The results of each query were sorted by the system specific rank value in descending order and the 1,000 highest-ranking records were collected to generate the result submission file. For numerous queries the system retrieved less than 100 documents and in some cases nearly no documents at all.

Results

As pointed out previously, our results were disappointing. Out of the 33 topics that had relevant documents, the Dutch-English multilingual run only retrieved relevant documents for approximately 70% (23) of them. The English monolingual run did slightly better retrieving relevant documents for approximately 76% (25). We believe that the low number of relevant documents for a large number of

topics in the test collection has affected the average precision measure (see *Analysis*) and therefore report the following numbers with some reservation. Average precision is 0.0364 for our cross-lingual run and 0.0678 for our monolingual run. A recall-precision table will not be presented since we would have to change the scale to make it show anything meaningful. As well, the graph will not provide a fair representation. Our Boolean system failed to retrieve the full 1000 documents for a large number of queries (we retrieved a total of 24,571 documents out of a possible 33,000 for cross-lingual and 15,057 out of 33,000 for monolingual).

In an effort to determine whether the problems we encountered were system based, we ran the identical set of queries on the Mirror DBMS system. The Mirror DBMS system combines information retrieval and data retrieval and uses statistical language models for information retrieval [9, 10]. The results improved drastically. For the cross-lingual run average precision improved by about 228% (new average precision 0.1197). For the monolingual run average precision improved by about 435% (new average precision 0.3630) (see figure 2). Interestingly, the monolingual results had a much larger improvement.

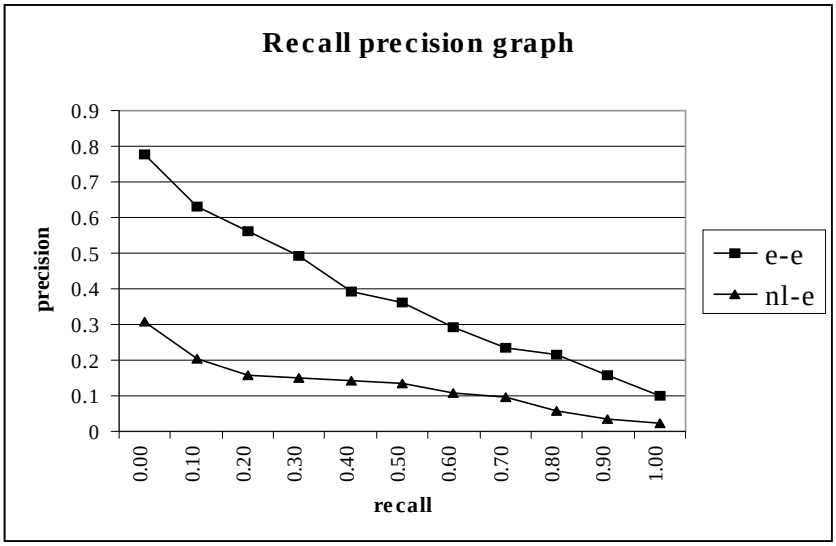

Fig. 2. Interpolated recall-precision using the Mirror DBMS

Analysis

The original results cannot just be blamed on the fact that most of the translations had to be removed to reduce the length of the queries (see *Experimental Setup*). Clearly, our monolingual results are also disappointing. We speculate that the lack of sophisticated linguistic processing, and techniques such as query expansion are reasons for our disappointing results. It is important to realize that the main reason for these results is the unsatisfactory retrieval capability of the commercial relational database used in the initial experiments. Additional experiments using the Mirror DBMS system show enormous performance improvements.

There are, however, issues regarding the test collection used in these experiments that impacts the evaluation of the results. Many of the topics only have a very limited number of relevant documents. Out of 40 topics, 7 topics do not have any relevant documents and these topics were left out of the analysis. This left 33 topics. Out of 33 topics 33% (11 documents) of documents have fewer than 10 relevant documents. And 18% of those (33 documents) have 5 or fewer relevant documents. The lack of relevant documents is problematic for measures such as average precision because averages are sensitive to large differences between numbers [11]. Topics 4 and 30, for example, only have 1 relevant document each. If this document is retrieved on rank 1 precision is 1 but if it is retrieved at rank 2 precision drops to 0.5. Average precision is also very sensitive to queries that perform poorly and these are represented in greater abundance in CLIR where extra noise is added in the translation. To soften

the impact of bad queries, a test collection should provide a larger number of topics to reduce the effect these queries might have. 33 topics alone might not be enough.

The shortage of relevant documents also affects precision (X) measures. Hull [12] suggests using high precision measures for cross-language system evaluation because they best reflect the nature of CLIR. In an ad hoc cross-lingual search, users are less likely to go through large numbers of documents to assess their relevance since they are not likely to be proficient in the language. It is important therefore to rank relevant documents at a high level. In addition, cross-lingual searches tend to benefit substantially from relevance feedback since this adds new foreign language terminology to the query that might be lacking in the original search. Here too it is important to rank relevant documents highly. Precision (10) is a good indicator of a system's ability to rank relevant documents highly. The problem with this test collection is that for 33% of the topics, a system could never have a perfect precision (10) score even if a system managed to retrieve all the relevant documents in the top 10.

Future Work

After a more careful analysis of the results described in this paper we plan on carrying out system testing exploring the system features more carefully. We plan on examining the translation from the query to the logical representation and the incorporation of query expansion and automatic relevance feedback.

Acknowledgements

The researchers would like to thank Arjen de Vries (CWI and University of Twente) for running our queries on the Mirror DBMS and for providing us with the retrieval results.

References

1. Oard, D. and Diekema, A.: Cross-Language Information Retrieval. In: Williams, M. (ed.): Annual Review of Information Science (ARIST), Vol. 33. Information Today Inc., Medford, NJ, (1998) 223-256
2. Ballesteros, L. and Croft, B.: Dictionary Methods for Cross-Lingual Information Retrieval. In: Proceedings of the 7th International DEXA Conference on Database and Expert Systems, September 9-13. Zürich, Switzerland. Springer-Verlag, New York, NY (1996) 791-801
3. Ballesteros, L. and Croft, B.: Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval. In: Proceedings of the Association for Computing Machinery Special Interest Group on Information Retrieval (ACM/SIGIR) 20th International Conference on Research and Development in Information Retrieval; 1997 July 25-31; Philadelphia, PA. ACM, New York, NY (1997) 84-91.
4. Oard, D. and Hackett, P.: Document Translation for Cross-Language Text Retrieval at the University of Maryland. In: Proceedings of the 6th Text REtrieval Conference (TREC-6); 1997 November 19-21. National Institute of Standards and Technology (NIST), Gaithersburg, MD. (1998) 687-696
5. Kraaij, W.: Multilingual Functionality in the Twenty-One Project. In: American Association for Artificial Intelligence (AAAI) Symposium on Cross-Language Text and Speech Retrieval; 1997 March 24-26; Palo Alto, CA. (1997) 127-132
6. Dumais, S. T.; Letsche, T. A.; Littman, M. L.; and Landauer, T. K.: Automatic Cross-Language Retrieval Using Latent Semantic Indexing. In: American Association for Artificial Intelligence (AAAI) Symposium on Cross-Language Text and Speech Retrieval; March 24-26; Palo Alto, CA. (1997) 15-21
7. Buckley, C.; Mitra, M.; Walz, J.; and Cardie, C.: Using Clustering and Super Concepts within SMART: TREC 6. In: Proceedings of the 6th Text REtrieval Conference (TREC-6); November 19-21; National Institute of Standards and Technology (NIST), Gaithersburg, MD. (1997) 107-124
8. Extensions to SQL Server to Support Full-Text Search
http:/www.microsoft.com/technet/SQL/Technotes/sq17fts.asp
9. de Vries, A. and Hiemstra, D.: The Mirror DBMS at TREC. In: Proceedings of the 8th Text Retrieval Conference. TREC-8, NIST Special Publication 500-246. National Institute of Standards and Technology (NIST), Gaithersburg, MD. (2000) 725-734
10. de Vries, A. and Hiemstra, D.: Relating the New Language Models of Information Retrieval to Traditional Retrieval Models. CTIT Technical Report TR-CTIT-00-09, (2000)
http://wwwhome.cs.utwente.nl/~hiemstra/papers/tr-ctit-00-09.ps
11. Buckley, C. and Voorhees, E.: Theory and Practice in Text Retrieval System Evaluation. A Tutorial Presented in Conjunction with the 22nd Annual International ACM SIGIR Conference on Information Retrieval. Berkeley, CA.. August 15, 1999. ACM, New York, NY (1999)

12. Hull, D. A.: Using Structured Queries for Disambiguation in Cross-Language Information Retrieval. In: American Association for Artificial Intelligence (AAAI) Symposium on Cross-Language Text and Speech Retrieval; March 24-26; Palo Alto, CA. (1997) 84-98